# Quantification of Visual Capabilities using Augmented Reality Displays

Mark A. Livingston*

3D Virtual and Mixed Environments Laboratory
Naval Research Laboratory

## Abstract

In order to be able to perceive and recognize objects or surface properties of objects, one must be able to resolve the features. These perceptual tasks can be difficult for both graphical representations and real objects in augmented reality (AR) displays. This paper presents the results of objective measurements and two user studies. The first evaluation explores visual acuity and contrast sensitivity; the second explores color perception. Both experiments test users' capabilities with their natural vision against their capabilities using commercially-available AR displays. The limited graphical resolution, reduced brightness, and uncontrollable visual context of the merged environment demonstrably reduce users' visual capabilities. The paper concludes by discussing the implications for display design and AR applications, as well as outlining possible extensions to the current studies.

**CR Categories:** I.3.7 [Computer Graphics]: Three-dimensional Graphics and Realism—Virtual Reality; H.5.2 [Information Interfaces and Presentation]: User Interfaces—Evaluation/methodology; H.1.2 [Models and Principles]: User/Machine Systems—Human Factors;

**Keywords:** augmented reality, visual acuity, contrast sensitivity, color perception

## 1 Introduction

Resolving objects and perceiving colors are fundamental tasks of human vision. Performance of these basic functions is difficult using optical see-through augmented reality (AR) displays. AR systems require graphical designs that enable users to discern graphical cues. In order to complete cognitive tasks such as recognizing familiar objects and reacting to visual information, users must perceive the graphical cues as the system designer intended. Otherwise, they will be unable to perform higher-level tasks under either experimental or operating conditions. Thus tests of perceptual capabilities also inform the results of application-level or cognitive tasks.

There are a few potential sources of the difficulty in resolving objects and perceiving colors in AR displays: effects of the optical elements within the display, effects of the display device, and effects due to the surrounding visual context. The first and last affect the user's perception of both the graphics and the real world; the second affects only the presentation of the graphics. The experiments discussed here do not directly investigate the sources; however, they do introduce viewing conditions that elucidate them.
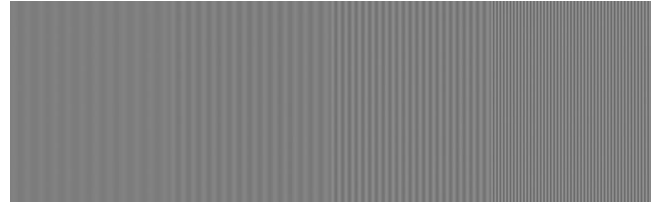
---

*E-mail: mark.livingston@nrl.navy.mil

Figure 1: A sine wave that increases contrast but decreases the visual angle of the pattern (increases the frequency) from left to right.

### 1.1 Visual Perception Measures

The quantity most people consider when they talk about visual capability is *visual acuity*, the ability of the observer to discriminate fine details in the visual field. The measure of visual acuity is the smallest stimulus that the observer can resolve. Normal visual acuity is approximately one minute of arc at a distance of 20 feet [18]. The most common type of test used to measure this quantity is the Snellen visual acuity chart. One major problem with this test is that the letters may vary in their perceptual difficulty. For example, an 'L' is easier to perceive than 'E' at the same size.

*Contrast sensitivity* describes the observer's ability to discern differences in the luminance values across an image. This has been accepted as part of a comprehensive approach to describing visual capabilities. Contrast is frequently expressed by the Michelson definition:

$$C = \frac{L_{\max} - L_{\min}}{L_{\max} + L_{\min}},$$

where $L_{\max}$ and $L_{\min}$ are, respectively, the maximum and minimum luminances in the image. The contrast in the image influences the observer's visual acuity score; at higher levels of contrast, the human eye is capable of detecting smaller details. Sine-wave gratings, such as in Figure 1, provide a convenient way to test the combined result of an observer's visual acuity and contrast sensitivity [8].

*Color perception* results from a complex set of retinal responses to light. The three types of cones (red, green, and blue) in the retina respond to different wavelengths of light, creating the effect people interpret as color. The Commission Internationale de l'Éclairage (CIE) defined three standard primaries to describe color in 1931, leading to the CIE chromaticity diagram (Figure 2) [4]. Color names may be ascribed to regions of the graph [10]; however, perception of color is more complicated than the diagram captures; for example, the color called purple is seen along several lines within the space. This space is not perceptually uniform; that is, the distance in this color space is not perceived as being equal when traveled in one direction as when traveled in another direction.

There are, however, color names that are consistently used. Smallman and Boynton [19] identified eleven basic color terms; of these, eight are chromatic and have one-word English names: red, green, blue, yellow, purple, orange,

| | | Form Approved |
|---|---|---|
| **Report Documentation Page** | | OMB No. 0704-0188 |

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

| 1. REPORT DATE<br>**2006** | 2. REPORT TYPE | 3. DATES COVERED<br>**00-00-2006 to 00-00-2006** |
|---|---|---|
| 4. TITLE AND SUBTITLE<br>**Quantification of Visual Capabilities Using Augmented Reality Displays** | | 5a. CONTRACT NUMBER |
| | | 5b. GRANT NUMBER |
| | | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | | 5d. PROJECT NUMBER |
| | | 5e. TASK NUMBER |
| | | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)<br>**3D Virtual and Mixed Environments Laboratory,Naval Research Laboratory,Washington,DC,20375** | | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES
**Proceedings of International Symposium on Mixed and Augmented Reality, Santa Barbara, CA, 22--25 October 2006**

14. ABSTRACT

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT<br>**unclassified** | b. ABSTRACT<br>**unclassified** | c. THIS PAGE<br>**unclassified** | **Same as Report (SAR)** | **10** | |

**Standard Form 298 (Rev. 8-98)**
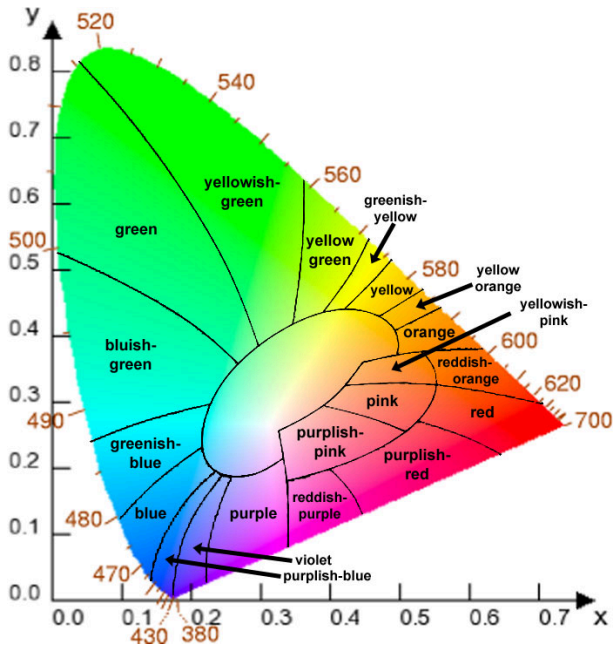Prescribed by ANSI Std Z39-18

Figure 2: The CIE 1931 chromaticity diagram with color names.

brown, and pink. (The achromatic names are black, gray, and white.) They found these colors to be maximally discriminable and unambiguously named, even cross-culturally.

## 1.2 Visual Perception in Head-worn Displays

It has long been known that low resolution in head-worn displays for virtual environments reduces the effective visual acuity of the user. An early informal experiment showed that a user wearing a VPL EyePhone Model 1 – resolution of $185 \times 139$ over an $80° \times 60°$ field of view (FOV) – could achieve a Snellen score of only approximately 20/250, which would be legally blind [9].

VEPAB [12] included Snellen eye charts for measuring the user's visual acuity. VEPAB tests reported that 24 users needed a mean distance to the chart of 4.65 ft to read the top line of a standard Snellen chart, which equates to a score of 20/860. This was achieved in a Virtual Research Flight Helmet display, which had a resolution of $238 \times 234$ in a $50° \times 41°$ FOV for each eye and should have enabled a Snellen score of 20/250. Color vision tests indicated that all users had normal color vision, but the last eight participants did not achieve perfect color vision scores in the virtual environment, leading to a hypothesis that the display may have been changing color character over time.

Video mixing AR systems limit the user to the resolution (spatial and color) of the camera, modulated by the display quality. Visual acuity, color perception, and hand-eye coordination with the real world have been tested in such systems [5]. Visual acuity through the camera was degraded, but no quantitative data are reported. Success rate on a Dvorine pseudo-isochromatic color test for color blindness dropped from 97.3% to 91.3%, remained at that level during testing, and rose to 96.7% in a post-test. Color identification dropped from 98.9% accuracy to 62.2% accuracy. Some adaptation occurred; after completion of the experimental task, color identification rose to 70.0% accuracy while still wearing the AR display. Accurate (100.0%) color perception returned after removing the display. No details were given on what constituted accuracy in color perception.

A test of four optical see-through AR displays [20] investigated the smallest real targets visible from one meter with the display off and with the display showing a blank screen. The latter condition implies that the display emits some light and, in the case the Sony Glasstron PLM-50, enables a filter that reduces transmittance of the light entering from the environment. Two binocular displays showed differences in these two conditions. The Glasstron (33° measured horizontal FOV) allowed 1 mm targets with no power (filter off) but only 6 mm targets with power (and filter) on, and I-glasses (25°) 0.5 mm and 3 mm. MicroOptical Corp. Clip-On (10°) and EyeGlass (17°) both allowed users to see 0.5 mm targets.

ARPAB [11] adapted VEPAB to AR displays. Testing of 20 subjects showed that a Sony Glasstron (SVGA, 27° horizontal FOV) yielded 20/40 Snellen scores; a Microvision Nomad (SVGA, ≈ 20° horizontal FOV) yielded 20/30 and 20/40 scores. Another Sony Glasstron [14] (LDI-D100B) caused eight users with normal or corrected-to-normal vision (i.e. 20/20 or better) to drop at least one step in Snellen score (≈ 20/30) looking through the optics of the display at the same real-world target. All users were scored at 20/30 looking at a graphical chart. In a realistic AR setting, the background and the drawing style (color, surrounding graphical field) have an effect on the legibility of text [7].

Evaluation of a head-mounted projection display (HMPD) was done with a modified Landolt-C acuity test [6]. This test asked users to identify the direction of an opening in a square (up, down, left, or right) under three levels of light. The study found that the resolution of the display limited subjects to a resolution of 4.1 arcminutes, or a Snellen score of 20/82 for all lighting levels. The type of retro-reflective material needed for the HMPD affected performance with low contrast targets.

## 1.3 Experimental Apparatus

The experiments described here considered optical see-through displays. A Sony Glasstron LDI-D100B and a Microvision Nomad were tested against the user's natural (or corrected) vision via an SGI GDM-FW9011 monitor. The Glasstron uses two color LCDs and optics that yield a fixed focus distance of 1.2 meters. It may be used in binocular mode; however, these experiments used bi-ocular mode. (Both eyes receive the same image.) This and the fixed distance for all viewing should have avoided problems that may occur in binocular displays in applications that require changing focus [15]. For looking at Glasstron graphics, its hardware-controlled opacity setting was set to the lowest level. This is typical for usage of the display in indoor environments. For looking through the Glasstron, the device (and thus the filter) were off; this makes the filter more transparent than its lowest setting when on. The Nomad is a monocular, monochromatic (red) retinal scanning display with an adjustable focal distance. Its focal distance is controlled by a hardware slider which is not labeled by distance. The experimenter matched the focal distance to that of the Glasstron; it remained fixed in place for all users. The CRT was placed with its flat screen at 1.2 meters from the user's eye position. Each user placed his or her chin on a chin rest in order to fix the eyes' positions (Figure 3).

In order to fairly test visual acuity, patterns in the stimuli must have the same visual frequency, whatever the resolution of the display device. The stimuli for this experiment occupied the same size in the horizontal FOV of the user across all three display devices (Table 1). The horizontal FOV for each display was measured. For the monitor, this required measuring the physical screen size and computing

Figure 3: The experimental scene, consisting of the chin rest and monitor, with the Glasstron positioned for testing. This image corresponds to the viewing condition "thru_Glass," explained in Section 2.

| Display | Resolution | FOV | Image Size |
|---|---|---|---|
| Monitor | $1600 \times 1200$ | $22.6° \times 14.6°$ | $1552 \times 1164$ |
| Glasstron | $800 \times 600$ | $28.1° \times 20.6°$ | $624 \times 468$ |
| Nomad | $800 \times 600$ | $23.7° \times 17.6°$ | $740 \times 555$ |

Table 1: Measurements and desired sizes for stimuli in each display.

the angle. For the head-worn displays, the virtual image size at the focus distance of 1.2 meters was measured by having an observer direct the placement of markers along a measuring stick so that they were at opposite edges of the FOV of the graphical field within the see-through area of the display. This gave the virtual image size at 1.2 m; from this the FOV was computed. There was no tracking of the display in these experiments; the display was affixed to the chin rest and could not move.

## 2  Acuity and Contrast Evaluation

The evaluation used five viewing conditions.

1. **realVision**, viewing of monitor with no head-worn display – subjects' natural or corrected vision (binocular)

2. **thru_Glass**, viewing of monitor through the Glasstron, with the display and the filter off (binocular)

3. **GlassWhite**, Glasstron graphics with a white background and filter at its most transmissive (bi-ocular)

4. **thru_Nomad**, viewing of monitor through the Nomad, with no graphics displayed (monocular)

5. **NomadWhite**, Nomad graphics with a white background (monocular)

| Display | Pixels per Degree (hor) | Minutes per Pixel | Snellen score |
|---|---|---|---|
| Monitor | 70.9 | 0.85 | 20/20 |
| Glasstron | 28.5 | 2.11 | 20/40 |
| Nomad | 33.8 | 1.78 | 20/30 |

Table 2: Conversion of resolution and horizontal FOV (from Table 1) to approximate expected Snellen score based on normal visual acuity.

### 2.1  Objective Measurements

One may evaluate the visual acuity possible with a particular viewing condition through trigonometry. These computations guided the hypotheses for the human subject evaluations. They also provided insights to the impediments caused by the displays and the human visual system's ability to adapt to them. Table 2 shows the measurements of each display computed from the horizontal resolution. The vertical resolution (computable from Table 1) was slightly better for each display, but not by enough to significantly change the expected Snellen score.

The CIE xyY coordinates of the gray values in the stimuli were measured with an LMT C1200 Colormeter under the viewing conditions described above. For the Glasstron and Nomad conditions, the colormeter was aimed through the optics of the display device. Unfortunately, the colormeter's aperture has a greater area than the Glasstron and Nomad displays. These measurements were taken in the standard office lighting conditions (fluorescent overhead fixtures) at the 1.2-meter viewing distance used in the experiments. Both the colormeter aperture and the lighting issues limit the accuracy of contrast measurements for the head-worn displays, as described in Section 4.

The Y value in the CIE color specification roughly equates to luminance, while the x and y coordinates give chrominance. Luminance values for the sine waves were measured by showing squares of gray shades at the sizes given in Table 1 and using only the Y value. Figure 4 shows the contrasts for the user study for each viewing condition, graphed against the theoretical values computed directly from the gray levels in the images. The theoretical values were chosen based on results from a pilot study: 0.0125, 0.0250, 0.0500, 0.1000, and 0.2000. These were not achieved perfectly in the digital imaging process, however. The maximum and minimum values for the image intensity were computed to be equidistant from the mean (128 in a range of $[0, 255]$) and achieve the desired contrast. The discretization of the image gray level perturbed the contrast from the desired values. For the lowest contrast, the range of intensities was $[126, 129]$ and for the highest contrast, the range was $[102, 153]$.

All displays narrowed the output luminance range and reduced the contrast. The Nomad, which was designed to be able to be used outdoors, demonstrated the greatest contrast, while the Glasstron, which was designed to be an alternate display for a mobile electronic devices, had the lowest contrast. Note that the presence of the graphics in the Glasstron made little difference in the contrast measurements. The thru_Nomad viewing condition tested at the highest contrast. The Nomad's graphics appeared to exhibit lower contrast than simply looking through the Nomad. This was counterintuitive and may have resulted from the difficult nature of measuring the contrast of the retinal scanning display; it may not reflect the relative order of contrast a user experienced. The Nomad also reduced the brightest luminance from the monitor less than the darkest luminance; this resulted in higher contrast in the thru_Nomad condition than in the realVision condition.
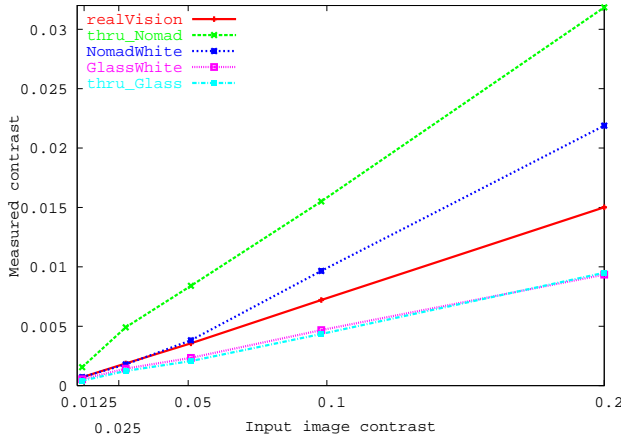
Figure 4: Contrast measurements for the viewing conditions used in the acuity and contrast experiment as a function of the theoretical contrast computed from the gray values used in the images.

## 2.2 Design of Perceptual Experiment

The perceptual user study used sine wave patterns (Figure 5) to measure the effect of the head-worn displays. The experimental design used four visual frequencies for the targets and five levels of Michelson contrast. The latter were chosen in order to encompass threshold contrast, meaning that users should have been perfectly accurate at the highest contrast and reduced to guessing at the lowest contrast.

Visual frequencies were chosen similarly. For each display, each frequency was converted to the number of pixels for the wavelength. The highest frequency was approximately 22% of the frequency perceived with "normal" visual acuity. However, during pilot testing, a higher frequency proved to be nearly impossible to discern at any level of contrast using the Glasstron. The highest frequency was 13.3 cycles per degree, or about three pixels at SVGA resolution in the AR displays, so that the pixel resolution was not the limiting factor in any viewing condition.

The stimuli were created by a program that computed the wave's amplitude at every pixel and computed a shade of gray based on the contrast chosen for that image. This per-pixel sampling of the wave's amplitude provided anti-aliasing of the rendered waves. The control software loaded one image for each of four frequencies, five contrast levels, and four orientations of the pattern, or 80 images for each viewing condition. Two extra frequencies (40 extra trials) were used for the realVision, thru_Glass, and thru_Nomad conditions to attempt to encompass threshold acuity for all frequencies and contrasts and generate data for Figure 6; these trials were not used in the statistical analysis.

The order of the three display devices used – not the viewing conditions – was counterbalanced with a Latin squares design. Thus the thru_Glass and GlassWhite viewing conditions were always consecutively experienced by subjects, but not necessarily in that order. This level of counterbalancing was done to reduce configuration changes needed in the environment (and thus the time needed from each subject). Since order effects were not expected, this was deemed an acceptable design.

### 2.2.1 Variables and Hypotheses

The experiment used the following four independent variables with the hypotheses noted.
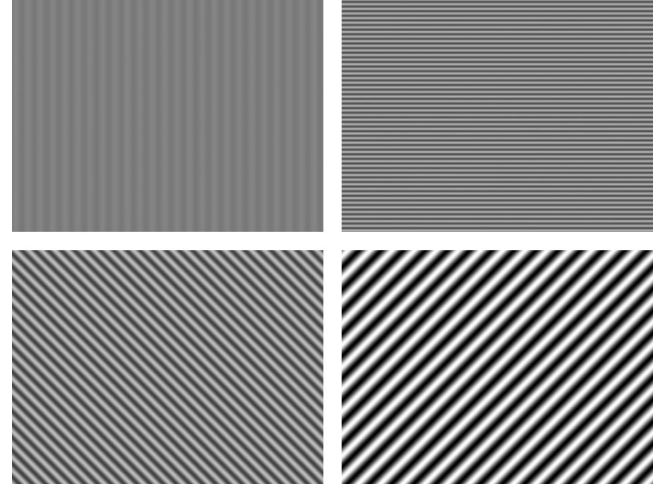


Figure 5: Four orientations of sine wave targets were used, and the frequency and contrast were varied. The frequency equates to a level of visual acuity. The contrast of the lower right target is greater than that used in the experiment, and is shown for illustrative purposes.

- *Visual frequency* ∈ { 13.3 Hz, 8 Hz, 4 Hz, 2.7 Hz }
  Hypotheses: Higher frequencies would yield lower performance in both accuracy and time (consistent with practical experience that smaller things are harder to see [18]); all users would be accurate for all frequencies in the realVision condition (as per the estimate in Table 2).

- *Contrast* ∈ { 0.0125, 0.0250, 0.0500, 0.1000, 0.2000 }
  (The five theoretical levels of contrast are maintained as labels for ease of reference.) Hypotheses: Lower contrast would yield slower, less accurate responses (consistent with practical experience that lower contrasts make it more difficult to discern boundaries [18]); users would perform better than guessing with their natural vision for even the lowest contrast.

- *Viewing condition* ∈ { realVision, thru_Glass, Glass-White, thru_Nomad, NomadWhite }
  Hypotheses: Based on experience with the displays, the Glasstron conditions would be the most difficult; users would be nearly flawless in the realVision condition.

- *Orientation* ∈ { horizontal, vertical, slanted, diagonal }
  Hypothesis: The "oblique effect" [1] predicts that observers should be approximately equally accurate with horizontal and vertical targets, and both of these should enable higher performance than diagonal or slanted targets (Figure 5). In some previous experiments with sine waves [8], users were most accurate with vertically-oriented targets, which implies a second possible hypothesis.

Each user thus completed $4 \times 5 \times 5 \times 4 = 400$ trials from which data was analyzed, plus 120 for the extra data for Figure 6.

### 2.2.2 Experimental Task

Users identified which of four orientations (Figure 5) the stimulus showed. Custom control software displayed the stimulus. For each stimulus, the user drew a line with the mouse that matched that orientation. When the user pressed the mouse button, the stimulus disappeared; only

the line the user was drawing was visible. When the user released the mouse button, the line's angle was computed, and the orientation (among the four choices) nearest to the line drawn was recorded. The program enforced a minimum line length; insufficient length resulted in the program warning the user and redisplaying the same stimulus. This occurred only seven times in 2400 total trials and was ignored in the analysis. Users were instructed to draw the line to within a reasonable approximation of their desired angle and that the software would compute the nearest to the four choices. They were instructed that there was a minimum line length of 20 pixels and that the program would ask them to redraw their response if they did not meet this threshold. They trained with five trials of the task on the first viewing condition they used. Halfway through each set and between sets, the user could take a break for as long as desired. One user took a one-hour break at one of these points to attend to work duties unrelated to this experiment.

The order of presentation of the targets was a random permutation computed at the time of the experiment. After the user drew the line to respond to a stimulus, the screen (monitor or head-worn display) rendered a black screen (which produced the see-through condition for the head-worn displays) for two seconds, then displayed the next stimulus. The control software recorded the response and the time the user viewed the stimulus (including multiple times in the case of too short a line being drawn).

### 2.2.3 Subjects

Five male and one female subjects between the ages of 33 and 52 completed the task. All self-reported normal or corrected-to-normal visual acuity. All were experienced computer users; three had significant experience with AR systems. All subjects volunteered and received no compensation. Since the Nomad is a monocular display, subjects were tested for their dominant eye. This test was performed as follows. The user formed a triangle with his or her hands; the thumbs overlapped to form the bottom side. The user was instructed to view a point (black tape on a light gray background) through this triangle with both eyes open. The user was then asked to close the left and right eye in sequence to determine which one of the eyes was viewing the point. The Nomad was placed in front of this eye and a black cloth in front of the other eye. One subject reported no dominance for either eye; this subject used the right eye.

### 2.3 Results

Figure 6 shows the measured contrast sensitivity function for each viewing condition using 62.5% accuracy as the threshold, which is standard for a four-alternative forced-choice task; this value gives consistency with statistical literature. The threshold value does not significantly change the contours or the trade-off between target size and contrast; the recommended threshold for optotypes is 70% or 75% [16]. Despite noisy data and extrapolation (to predict the lower bounds for how little contrast would enable acceptable user performance), the graph shows users were best in the realVision condition closely followed by the thru_Nomad and NomadWhite conditions. As expected, subjects struggled to perceive the targets in both thru_Glass and GlassWhite conditions. For frequencies greater than 8 Hz, subjects were reduced to guessing in the thru_Glass viewing condition. They were better in the GlassWhite condition but unable to perform the task at high frequencies and low contrasts. Table 3 gives summary statistics for the independent variables.
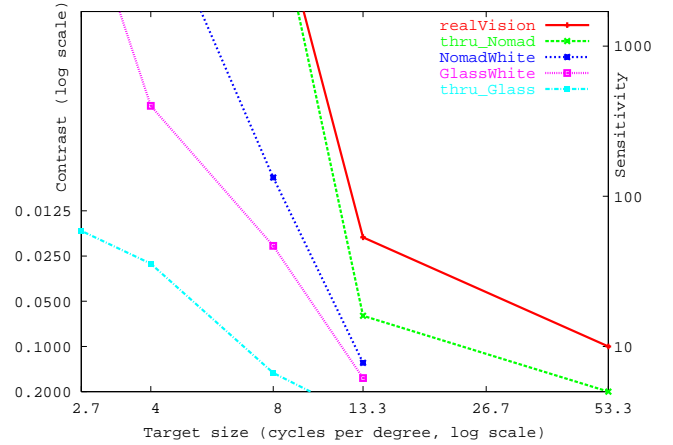


Figure 6: Contrast sensitivity function resulting from the acuity and contrast experiment. The contours show the frequency (in cycles per degree) and contrast combinations for which subjects were judged to be able to discern the stimulus orientation for each viewing condition.

| | Condition | N | % | Mean | SD | SE |
|---|---|---|---|---|---|---|
| **Display** | realVision | 480 | 96 | 0.071 | 0.353 | 0.016 |
| | thru_Nomad | 480 | 90 | 0.150 | 0.486 | 0.022 |
| | NomadWhite | 480 | 85 | 0.225 | 0.563 | 0.026 |
| | GlassWhite | 480 | 79 | 0.352 | 0.716 | 0.033 |
| | thru_Glass | 480 | 57 | 0.715 | 0.881 | 0.040 |
| **Frequency** | 2.7 Hz | 600 | 96 | 0.062 | 0.339 | 0.014 |
| | 4 Hz | 600 | 93 | 0.093 | 0.398 | 0.016 |
| | 8 Hz | 600 | 76 | 0.318 | 0.684 | 0.028 |
| | 13.3 Hz | 600 | 42 | 0.737 | 0.863 | 0.035 |
| **Contrast** | 0.2000 | 480 | 93 | 0.115 | 0.435 | 0.020 |
| | 0.1000 | 480 | 88 | 0.194 | 0.550 | 0.025 |
| | 0.0500 | 480 | 83 | 0.271 | 0.625 | 0.029 |
| | 0.0250 | 480 | 78 | 0.369 | 0.722 | 0.033 |
| | 0.0125 | 480 | 66 | 0.565 | 0.835 | 0.038 |
| **Orient.** | horizontal | 600 | 83 | 0.197 | 0.528 | 0.022 |
| | vertical | 600 | 81 | 0.230 | 0.572 | 0.023 |
| | slanted | 600 | 72 | 0.390 | 0.752 | 0.031 |
| | diagonal | 600 | 72 | 0.393 | 0.757 | 0.031 |

Table 3: Summary statistics in acuity/contrast experiment. N is number of trials; % is percent correct over all users; Mean is mean error for a value; SD is standard deviation; SE is standard error.

### 2.3.1 Main Effects and Interactions

A $4 \times 5 \times 5 \times 4$ within-subjects analysis of variance (performed with the |STAT analysis package [17]) revealed statistically significant effects. Performing statistical evaluations requires an error metric for the discrete, forced-choice task. Certainly, a correct choice was zero error. The analysis defined a response of horizontal for a vertical stimulus to be an error of one, on the theory that the $90°$ turn was more perceptually similar in the pixelated grid of the displays than a $45°$ turn. Following this logic, interchanging the diagonal orientations (Figure 5, bottom) was an error of one, whereas interchanging a diagonal for either horizontal or vertical was an error of two. One could also define the error by the number of $45°$ turns required to reach the correct orientation from the response, which also yields an error of zero, one, or two. The results which follow were significant for both definitions, with one exception discussed last. Numbers given are for the definition using exchanges, not turns.
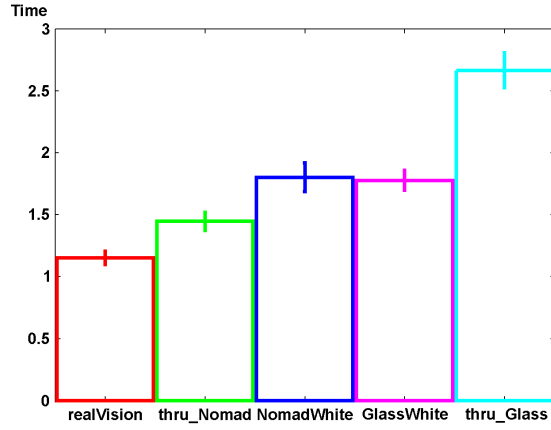
Figure 7: The effect of viewing condition on time does not follow the ordering of performance with viewing conditions in Figures 8 or 9.

The viewing condition ($F(4, 20) = 89.25$, $p < 0.001$), visual frequency ($F(3, 15) = 169.4$, $p < 0.001$), and contrast ($F(4, 20) = 50.2$, $p < 0.001$) have expected main effects on the error. All three showed a significant effect on response time; higher frequencies ($F(3, 15) = 22.2$, $p < 0.001$) and lower contrast ($F(4, 20) = 18.8$, $p < 0.001$) slowed users, as expected. The significant main effect of viewing condition on time ($F(4, 20) = 6.0$, $p \approx 0.002$) shows realVision was faster and thru_Glass was slower, as expected (Figure 7). But thru_Nomad was faster than the displayed graphics (NomadWhite and GlassWhite); this contrasted with the fact that in most respects, the NomadWhite viewing condition was nearly equivalent to the thru_Nomad condition.

The significant interactions give insight into what happened. Figure 8 shows the interaction between viewing condition and visual frequency ($F(12, 60) = 16.1$, $p < 0.001$). The graph shows the progression of viewing conditions, from realVision enabling the best performance to the Nomad and then the Glasstron. This difference largely disappeared for lower frequencies for most viewing conditions, emphasizing the difficulty the users had seeing their physical surroundings when looking through the Glasstron. But it also shows the progressively difficult nature of seeing small details with all of the viewing conditions, in roughly the order expected in the hypothesis for viewing conditions. Similarly, there is an interaction between the viewing condition and the contrast ($F(16, 80) = 3.2$, $p < 0.001$). The effect of contrast, however, separates the viewing conditions at the higher contrasts used in the study (Figure 9). At the highest frequency, users exceeded threshold contrast. At the lower frequencies, users always performed better than chance. The contrast sensitivity function (Figure 6) shows under what conditions the combined effect of frequency and contrast allows the user to discern the stimuli. Note that the ordering of the viewing conditions by increasing error did not mimic the order of decreasing contrast in Figure 4. There was a significant three-way interaction between the viewing condition, visual frequency, and contrast ($F(48, 240) = 5.8$, $p < 0.001$). A favorable value in neither the visual frequency nor the contrast completely overcame unfavorable values in the other.

There was a main effect of stimulus orientation, but only for the error definition that treats the horizontal and vertical orientations as one class and the two diagonal orientations as another class ($F(3, 15) = 9.0$, $p \approx 0.001$). Users had lower errors with horizontal and vertical stimuli than with the two diagonal stimuli, an observation of the oblique effect. How-
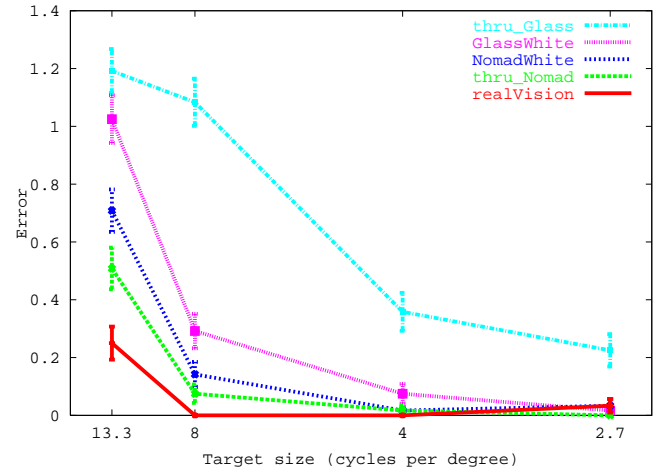


Figure 8: The interaction between the viewing condition and visual frequency shows how much more difficult it is to see the surrounding environment through the Glasstron. The error bars in this and all graphs show one unit of standard error.
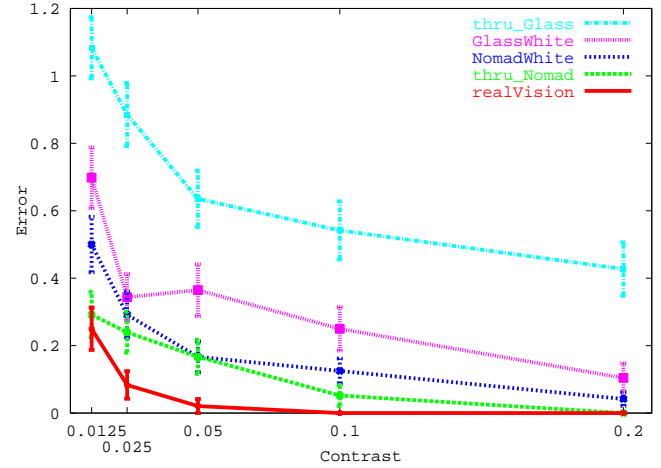


Figure 9: The interaction between viewing condition and contrast shows that the effect of contrast is not overcome quite as easily as the effect of the visual frequency.

ever, the error definition that uses turns does not show a significant effect ($F(3, 15) = 2.7$, $p \approx 0.083$). Some users reported using horizontal as a default guess, but it is not clear whether the choice of the definition of error should change the significance of the effect. Given that significance found with the first error definition matches the results found in similar experiments, this likely speaks more to the appropriateness of the error metric than an interesting result.

### 2.3.2 Afterimage errors

To check whether subjects had their visual palates cleansed between trials required the definition of an afterimage error. Two definitions were considered with two criteria each. Clearly, the first (shared) criterion is that response was incorrect. The second criterion is that the user's incorrect response matches something seen prior to the trial. One could say that the incorrect response should match the previous *stimulus*, or one could say that the incorrect response should match the previous *response*. That is, one could argue an afterimage is a physiological effect (matching what users saw)

|  |  | Total Errors | Physiological Afterimages | | Cognitive Afterimages | |
|---|---|---|---|---|---|---|
| **Display** | realVision | 20 | 7 | 35% | 8 | 40% |
| | thru_Nomad | 46 | 17 | 37% | 21 | 46% |
| | NomadWhite | 72 | 23 | 32% | 27 | 38% |
| | thru_Glass | 203 | 73 | 36% | 84 | 41% |
| | GlassWhite | 100 | 31 | 31% | 41 | 41% |
| **Frequency** | 13.3 Hz | 119 | 92 | 77% | 117 | 98% |
| | 8 Hz | 105 | 40 | 38% | 44 | 42% |
| | 4 Hz | 127 | 13 | 10% | 13 | 10% |
| | 2.7 Hz | 90 | 6 | 7% | 7 | 8% |
| **Contrast** | 0.0125 | 98 | 57 | 58% | 64 | 65% |
| | 0.0250 | 81 | 40 | 49% | 50 | 62% |
| | 0.0500 | 92 | 24 | 26% | 31 | 34% |
| | 0.1000 | 93 | 21 | 23% | 21 | 23% |
| | 0.2000 | 77 | 9 | 12% | 15 | 19% |
| **Orient.** | horizontal | 123 | 32 | 26% | 39 | 32% |
| | vertical | 107 | 30 | 28% | 33 | 31% |
| | slanted | 103 | 35 | 34% | 50 | 49% |
| | diagonal | 108 | 54 | 50% | 59 | 55% |

Table 4: The number of errors and potential afterimage errors for the independent variables. Examining the percentages reveals which significant effects were not reflections of the main effects.

or a cognitive effect (matching what users think they saw). Both definitions are analyzed here.

There were 447 errors out of 2400 trials; six of these occurred on the first stimulus within a set and thus are not afterimages. Of the 441 remaining errors, the physiological definition identified 151 potential afterimage errors. The cognitive definition found 181 potential afterimage errors. (The intersection of the two sets contained 123 errors.) Table 4 breaks the errors down for the independent variables.

There were significant differences between viewing conditions with both the physiological definition ($F(4, 20) = 25.2$, $p < 0.001$) and the cognitive definition ($F(4, 20) = 22.6$, $p < 0.001$). However, the percentage of afterimage errors for each viewing condition was nearly equal. This merely reflects the main effect of viewing condition on error.

There were significant differences for the visual frequency with both the physiological ($F(3, 15) = 79.7$, $p < 0.001$) and cognitive ($F(3, 15) = 114.3$, $p < 0.001$) definitions. The highest frequency (13.3 Hz) had the most errors, and the percentage of errors identified as potential afterimage errors was twice that of the second-highest frequency for both definitions. In fact, nearly all errors made at the highest frequency were identified as potential afterimage errors. This certainly merits further investigation. Thus one may conclude that the significant main effect of frequency shows that lower frequencies induce fewer afterimage errors for both definitions. The effect appeared to diminish by the second-smallest frequency chosen for the experiment.

There were also significant differences for the contrast for the physiological ($F(4, 20) = 12.1$, $p < 0.001$) and cognitive ($F(4, 20) = 10.1$, $p < 0.001$) definitions. The lowest contrast (0.0125) had the most errors. The percentage of errors identified as potential afterimage errors was approximately the same for the two lowest levels of contrast, and those were nearly twice the percentages for the next two levels of contrast. It is unclear whether the effect was diminishing with the contrast levels in this experiment, however.

The effect of stimulus orientation was unclear. The cognitive definition appeared to fit the hypotheses better than the physiological definition. However, the physiological def-

inition yielded a barely significant effect ($F(3, 15) = 3.4$, $p \approx 0.044$), whereas the cognitive definition did not reach significance ($F(3, 15) = 2.3$, $p \approx 0.125$). One could infer that the cognitive definition is more appropriate but the effect is not significant; this merits further investigation.

## 3   Color Evaluation

The second task moved toward the goal of measuring consistency of color perception between viewing conditions. This began by selecting a set of colors distributed in CIE color space; CIE xyY coordinates were measured with the C1200.

### 3.1   Objective Measurements

The goal of the objective measurements of this evaluation was to measure any color shift. However, the C1200 integrates over its circular aperture (diameter of three inches). This immediately presented a problem in trying to measure the chromaticity values. The exit pupil of both head-worn displays was much smaller than this aperture. While the surrounding region around the display's exit pupil was masked with black cloth, this still affected the measurement. Using a Hoffman Engineering TSP-90 photometer to measure the relative intensity of the graphics of the head-worn displays and the monitor and the areas of the exit pupil, one can compute a normalization factor. If the mask were not aligned properly, this would introduce an error into the measurement. The optics of the head-worn displays had a larger field of view than the colormeter; this gathering of incoming light also affected chromaticity measurements. Although the black cloth should have produced no reflections, it was not clear that this was true under the lighting conditions.

This normalization led to the plots of the samples for the four viewing conditions: realVision, thru_Glass, GlassWhite, and GlassBlack (explained in the next section). Figure 10 shows the four images, generated by taking chrominance (xy) values from the colormeter, normalizing them as above, then plotting them within CIE 1931 color space. While the overall scale of the plots was confounded by the normalization issue, the arrangement within the perimeter of the sampled area was due to the viewing conditions. This was, in part, an effect of the changing monitor gamut for the desktop monitor and the Glasstron.

### 3.2   Design of Perceptual Experiment

The second experiment was interleaved with the acuity and contrast experiment to reduce configuration changes necessary during the experiment. Users verbally identified the color of monochromatic stimuli from among the eight chromatic colors identified as maximally discriminable [19]; the experimenter entered a keystroke to record the response. The stimuli were the same size as the sine wave patterns given in Table 1. Users were instructed to pick exactly one of the eight choices. They were presented a palette and cautioned not to try to pick the closest hue from the palette, but simply to apply one of those eight names to the color they saw in the stimulus. Subjects again trained with five practice trials in only the first viewing condition they used (among the color displays). The same subjects completed this task. No subject reported any color-blindness. The set of color samples were again presented in a randomly permuted order. As these sets of trials were interleaved with the acuity and contrast experiment, the same opportunity for rest breaks occurred in this experiment.

The only variable maintained from the first task was the viewing condition. However, there were changes to the values. This test used the same Sony Glasstron with the same settings. A new viewing condition for the Glasstron was created by adding a black background (GlassBlack). This viewing condition was bi-ocular. Since the Nomad is monochromatic, it was not used in this test. The only new variable was the chromaticity of the presented color, expressed in Figure 10. The hypothesis was that the perceived colors would be inconsistently labeled near the boundaries between colors, and that there would be a shift between viewing conditions.

## 3.3 Results

The global trends in the shift of perceived color may be seen in Figure 10. The color brown was one interesting example; it expanded in the thru_Glass condition but disappeared in the GlassWhite condition. The bright background seemed to create a conflict with the dark color. Similarly, orange nearly disappeared in the GlassWhite condition. It was used most in the GlassBlack condition, when all colors appeared to be darker. Purple seemed to shrink in the GlassBlack condition. There was more overlap between the regions in each of the three Glasstron-based viewing conditions. (At least three subjects had to give a label for that sample to be in a region.) Color perception is highly dependent on context. The Glasstron tends to make all graphical objects significantly lighter because they are semi-transparent; when graphics are displayed, it tends to make the real background significantly darker with its opacity shutter. These clearly affected users' perceptions of colors.

An obvious error metric may be derived from the differences between the red, green, and blue components of the input color. Since there were only eight responses for the colors, this was really a discrete metric. Also, since there is really no objective basis for color names, the metric measured differences between the realVision condition and the other three conditions. That is, error was further defined as a change from the realVision condition, with no head-worn display intervening. The metric worked by considering the value of each component on a scale from 0.0 to 1.0. If the component was at 1.0 in a color but at 0.0 in another color, then there was a difference of 1.0 between those colors for that component. This metric was summed over all components. Colors such as orange and purple introduced differences between 0.0 and 1.0 to the metric. Table 5 gives the complete set of discretized error values. While this does not correspond to the perceptual spacing of the colors, it was a good approximation to the difference between the colors.

Another error metric is the angle around the center (white) point in the CIE chromaticity diagram. While this color space is not perceptually uniform, this metric is still a measure of the distance between two colors. The results of the analyses for the two error functions are similar.

There were significant differences in error for certain colors with the component-based ($F(59, 295) = 2.5$, $p < 0.001$) and angle-based ($F(59, 295) = 3.2$, $p < 0.001$) error metrics. Six stimuli were identified by both error functions as among the colors with the highest error. Near the brown region, a sample was unlabeled entirely in the thru_Glass condition because no three subjects gave it the same label. Other problematic colors occur at the intersection of the red and orange regions, the boundary between green and yellow, the convergence of orange, brown, and yellow, and the unusual protrusion of the pink region into the purple in the thru_Glass condition. Perhaps the most interesting of these, however, was the stimulus that changed between pink and blue. This was counterintuitive, as those two colors are not considered perceptually close. However, there were some colors near the white center that defied consistent labeling. The blue region, perhaps because of the low contribution of blue wavelengths to intensity, grew into the area on the opposite side of the center from the blue corner of the diagram. This result merits further investigation.

## 4 Discussion and Future Work

The basic result for the acuity and contrast was as expected: higher frequencies and lower contrast objects were progressively harder to see, and the difficulty increased significantly from natural vision to looking through the Glasstron. The surprise was that just looking through the optical elements of the Glasstron was more limiting than looking at its graphics; the contrast with which the real environment could be seen was extremely low in normal indoor conditions. We have a sampling of the contrast sensitivity function (Figure 6) that may guide for AR system designers. More data could refine the resolution of this graph. Users were not able to achieve performance that would correspond to the maximum Snellen score for which the stimuli could test, which was below that implied by the pixel resolution of the displays, so this is an important design consideration for these displays.

The shift in color perception was remarkable, and has significant implications for programs that use color as a key for the user. Dark colors tended to disappear perceptually with light backgrounds. Since many AR applications have little control over their backgrounds and optical see-through AR displays have limited intensity resolution versus the background, this is a key design consideration for applications that use optical see-through AR displays. How this might interact with a color-blind user is another issue to be explored. Additionally, perceived color differences change with the size of objects [2]. All the backgrounds used in the acuity and contrast experiment were white; different backgrounds may produce different perceptions, as was seen in the color evaluation. These chromatic interactions will affect practical AR applications.

The display devices did not offer everything that an application designer might want. Notably, the Glasstron did not allow adjustment of the distance between the eyes. Each user's inter-pupillary distance was measured, but no separation was provided. Two users observed a slight horizontal disparity between the two eyes while using the Glasstron. While this probably did not affect the color perception, it certainly could have made the acuity and contrast task more difficult. The Nomad had a bright spot in the middle of its display; three users remarked that they found this bright spot helpful in identifying the orientation of lines. (It was not clear that it was actually helpful, however.) Two users reported increased eye strain both after using the Glasstron and after using the Nomad. One user noted that having the bright light of the Nomad in only one eye (and virtually no light in the other) required a notable adjustment to return to normal vision. With only six subjects, no analysis has been done to determine whether this related to performance, but this is a consideration for future studies. Total time in all viewing conditions ranged between 40 and 70 minutes.

There are many variables in these experiments, most of which could be sampled at a wider range and at greater resolution. Most importantly, it should be possible to squeeze a higher visual frequency into the stimuli for the head-worn display. The wavelengths of the frequencies chosen were not close enough to units of whole pixels. This created alias-

ing in the stimuli (beyond that which could be mitigated by anti-aliasing techniques in the rendering). No subjects (even those with significant computer graphics backgrounds) reported noticing this effect. This is not likely to be a significant problem for the accuracy of identification, but it might have caused the subject to have to search through the image for a region without aliasing, slowing down the response. The pause or display between trials could be varied to explore afterimage effects more thoroughly.

A related consideration to the contrast variable are the lighting conditions. In all trials, two pair of standard four-bulb fluorescent lights illuminated the experimental room, a $2.4 \times 4.1$-meter office. The walls were white and a dry erase whiteboard was on one wall, parallel to the user's view direction during the trials. The door was closed; no external light source provided significant lighting to the room. Many optometric exams are performed in darkened rooms, such that only the stimulus (e.g. Snellen eye chart) is lit (and perhaps the desktop for the optometrist to read or write notes). The National Academy of Sciences recommends a minimum contrast of 0.85 for optometric testing [16]. Despite these recommendations, no standardization exists in clinical practice [3]. Similarly, there is little consistency of lighting conditions for AR applications. Future tests could include lighting as a variable.

Rivalry between the eyes and interference with the background degrade performance on a 2D search task [13]. The current study avoided these issues by using solid backgrounds and either showing the same image to both eyes (binocular for monitor viewing, bi-ocular for the Glasstron) or placing a blocker in front of one eye. However, in a practical system, the user will not discern graphics that create rivalry or interference.

Some variables could not be examined with this experimental design and subject pool. There is insufficient data to examine gender differences and subject experience with AR displays. The block structure of viewing conditions (always showing either thru_Glass followed by GlassWhite or GlassWhite followed by thru_Glass) and interleaving the acuity/contrast stimuli sets with color stimuli sets may have produced undesired order effects, though this seems unlikely.

AR displays still reduce users' visual capabilities but not nearly as much as early displays did. These experiments started quantifying the reduction users suffer when donning a head-worn AR display. Users were able to complete simple tasks, but not at the accuracy to which they are accustomed or should be able to achieve based on display resolution. Clearly, higher-resolution and brighter displays that are currently available would promise better results for the acuity and contrast experiment. It remains to quantify how much improvement comes from the increase in resolution. The question of how much acuity and contrast sensitivity is necessary for a given task also remains open. The proper balance of enabling the user to see the real world and perceive colors as intended is elusive. With these types of experiments, application designers may begin to quantify the changes. Hopefully, this will lead to a better understanding of the general questions about perception in AR systems.

## Acknowledgments

## References

[1] Stuart Appelle. Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals. *Psychological Bulletin*, 78(4):266–278, October 1972.

[2] Robert C. Carter and Ellen C. Carter. Color coding for rapid location of small symbols. *Color Research and Application*, 13(4):226–234, August 1988.

[3] Lynda Charters. Vision screening standardization recommended. *Ophthalmology Times*, 15 April 2006.

[4] CIE. *Commission Internationale de l'Eclairage Proceedings*. Cambridge University Press, Cambridge, England, 1931.

[5] Rudolph P. Darken, Joseph A. Sullivan, and Mark Lennerton. A chromakey augmented virtual environment for deployable training. In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC 2003)*, December 2003.

[6] Cali Fidopiastis, Christopher Fuhrman, Catherine Meyer, and Jannick Rolland. Methodology for the iterative evaluation of prototype head-mounted displays in virtual environments: Visual acuity metrics. *Presence: Teleporators and Virtual Environments*, 14(5):550–562, October 2005.

[7] Joseph L. Gabbard, J. Edward Swan II, Deborah Hix, Robert S. Schulman, John Lucas, and Divya Gupta. An empirical user-based study of text drawing styles and outdoor background textures for augmented reality. In *IEEE Virtual Reality*, pages 11–18, March 2005.

[8] Arthur P. Ginsburg and William R. Hendee. *Quantification of Visual Capability*, pages 52–71. Springer-Verlag, 1992.

[9] Richard Holloway, Henry Fuchs, and Warren Robinett. *Virtual Worlds Research at the University of North Carolina at Chapel Hill as of February 1992*, pages 109–128. Springer-Verlag, 1992.

[10] Deane B. Judd. Colorimetry. Technical Report Circular 478, National Bureau of Standards, 1950.

[11] Sonny E.H. Kirkley, Jr. *Augmented Reality Performance Assessment Battery (ARPAB): Object Recognition, Distance Estimation and Size Estimation Using Optical See-through Head-worn Displays*. PhD thesis, Department of Instructional Systems Technology, Indiana University, May 2003.

[12] Donald R. Lampton, Bruce W. Knerr, Stephen L. Goldberg, James P. Bliss, J. Michael Moshell, and Brian S. Blau. The virtual environment performance assessment battery (VEPAB): Development and evaluation. *Presence: Teleporators and Virtual Environments*, 3(2):145–157, Spring 1994.

[13] Robert S. Laramee and Colin Ware. Rivalry and interference with a head-mounted display. *ACM Trans. on Computer-Human Interaction*, 9(3):238–251, September 2002.

[14] Mark A. Livingston, Catherine A. Zanbaka, J. Edward Swan II, and Harvey S. Smallman. Objective measures for the effectiveness of augmented reality. In *IEEE Virtual Reality 2005 (Poster Session)*, pages 287–288, March 2005.

[15] Mark Mon-Williams and John P. Wann. Binocular virtual reality displays: When problems do and don't occur. *Human Factors*, 40(1):42–49, March 1998.

[16] National Academy of Sciences. Recommended standard procedures for the clinical measurement and specification of visual acuity. *Advances in Ophthalmology*, 41:103–148, 1980.

[17] Gary Perlman. Data analysis in the unix environment. In *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface*, pages 130–138. Springer-Verlag, July 1982.

[18] Lorrin A. Riggs. Visual acuity. In *Vision and Visual Perception*, pages 321–349. John Wiley and Sons, 1965.

[19] Harvey S. Smallman and Robert M. Boynton. On the usefulness of basic colour coding in an information display. *Displays*, 14(3):158–165, 1993.

[20] Russell L. Woods, Ivonne Fetchenheuer, Fernando Vargas-Martin, and Eli Peli. The impact of non-immersive head-mounted displays (HMDs) on the visual field. *Journal of the Society for Information Display*, 11(1):191–198, 2003.
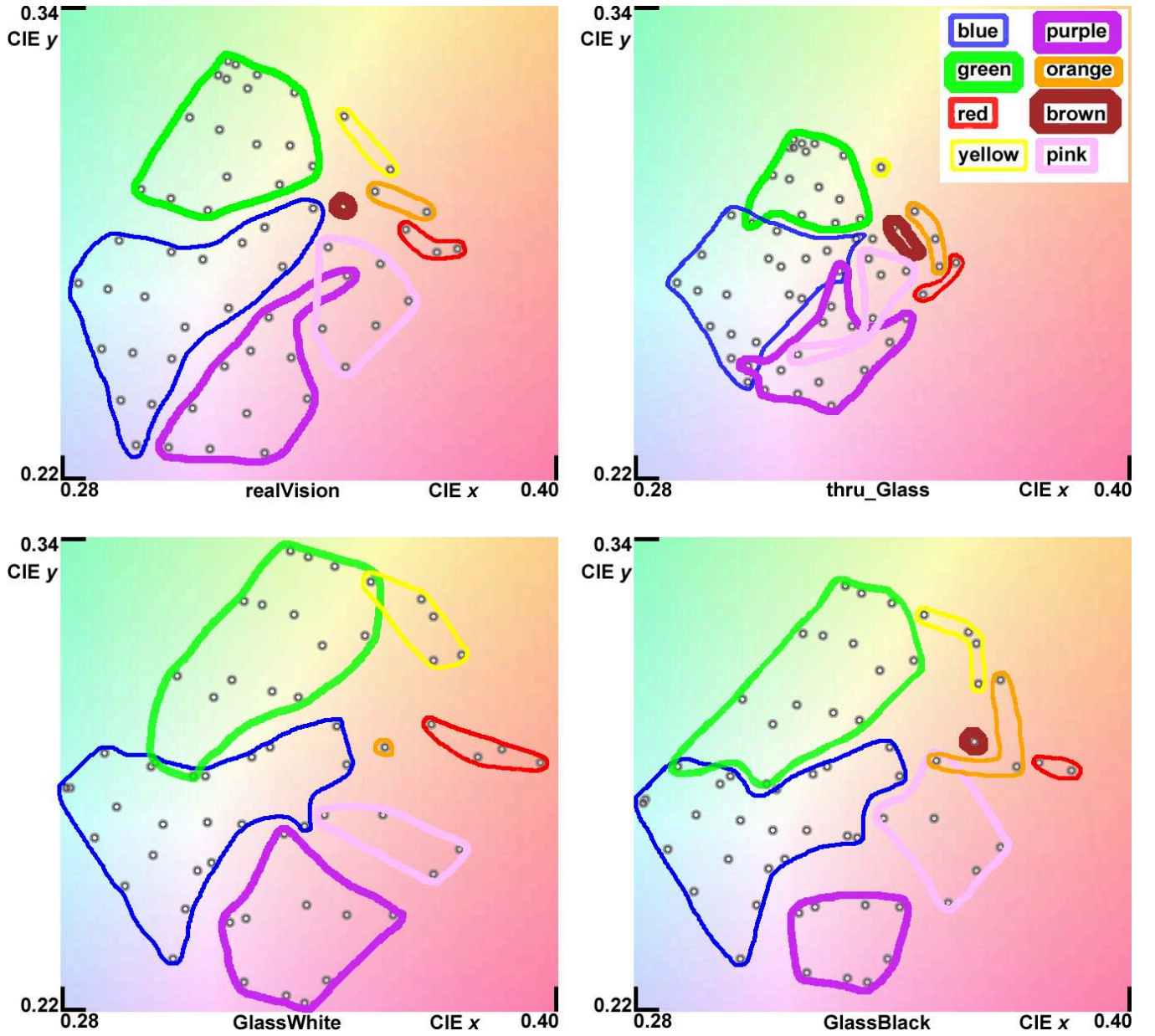
Figure 10: Cropped CIE color space with the labeled colors from the four viewing conditions. As indicated in Figure 2, green is in the upper left, blue in the lower left, red in the lower right, and yellow and orange in the upper right. The legend in the upper right of the thru_Glass condition applies to all four graphs. The arrangement of the color samples warps between viewing conditions, even discounting the scale, which is in turn confounded by the normalization problem. The samples had to be labeled by three users as a particular color for the plot to reflect that sample as that color. Note that brown disappears in the GlassWhite condition. Also note that in the thru_Glass condition, a sample was not labeled the same color by any three users; this appears between the brown and blue regions.

| Name | R | G | B | red | green | blue | yellow | purple | brown | pink | orange |
|---|---|---|---|---|---|---|---|---|---|---|---|
| red | 1.0 | 0.0 | 0.0 | 0.0 | | | | | | | |
| green | 0.0 | 1.0 | 0.0 | 2.0 | 0.0 | | | | | | |
| blue | 0.0 | 0.0 | 1.0 | 2.0 | 2.0 | 0.0 | | | | | |
| yellow | 1.0 | 1.0 | 0.0 | 1.0 | 1.0 | 3.0 | 0.0 | | | | |
| purple | 0.6 | 0.1 | 1.0 | 1.5 | 2.5 | 0.7 | 2.3 | 0.0 | | | |
| brown | 0.6 | 0.2 | 0.2 | 0.8 | 1.6 | 1.6 | 1.4 | 0.9 | 0.0 | | |
| pink | 1.0 | 0.8 | 1.0 | 1.8 | 2.2 | 1.8 | 1.2 | 1.1 | 1.8 | 0.0 | |
| orange | 1.0 | 0.6 | 0.0 | 0.6 | 1.4 | 2.6 | 0.4 | 1.9 | 1.0 | 1.2 | 0.0 |

Table 5: Table of the symmetric RGB-component-based error function for color labeling.